DOCUMENT RESUME

ED 093 931 TM 003 749

AUTHOR Bar-On, Ehud; And Others
TITLE The Use of Computers in Evaluating Teacher Competency.
PUB DATE Apr 74
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Computer Programs; Lesson Observation Criteria; Measurement Techniques; *Performance Based Teacher Education; Statistical Analysis; *Student Teachers; *Teacher Evaluation

ABSTRACT

Sixty-five student-teachers' performance was tested to determine laboratory grades. Grading was based on Category Observation System TDS. Since TDS categories are structures of two ordered facts where order has the same meaning, general score of pupil stimulation was computer calculated. Students' previous awareness as to score calculation enabled lesson planning to achieve the highest possible grade. The resulting score was highly correlated with supervisors' general evaluation and was preferred by students. Comparison of test and last lesson before receiving instruction showed dramatic improvement. Planned and actual lessons were compared regarding realistic planners, frequent categories and teaching sequences. (Author)

THE USE OF COMPUTERS IN EVALUATING TEACHER COMPETENCY

Ehud Bar-On and Rachel Levin

Technion - Israel Institute of Technology, Haifa, I S R A E L.

FIRST DRAFT

April, 1974 .

THE USE OF COMPUTERS IN EVALUATING TEACHER COMPETENCY

The title of this article seems to be a contradiction: "evaluation consists of an assessment of merit" (Scriven, Michael 1956) and a computer is not expected to value. In other words, one does not expect a machine to perform in a level of "thinking" which can be classified as "evaluation" according to Bloom's "Taxonomy". However, it was not intended to place on the computer the judgement of teacher performance during the test lesson. It has been attempted to build the authors' value scale, regarding teacher performance into the computer program. By assigning different weights to different classroom behaviors and by defining these classroom activities explicitly we intended to form a grading procedure which would reflect our values and yet will enable rigorous evaluation of the quality of instructional activities. Writing the grading procedure in such a way that it can be programmed, insures the rigorousness of the process. But there are still three problems to be solved:

1. How to transduce what is occuring in the lesson to numerical data available for the computer?
2. Which classroom activities have to be chosen?
3. Which weight should be assigned to each of the chosen behaviors?

The answer to the first question is not new: a trained observer using a category observation system and a time sampling method can convert what is occurring in the class to code numbers. To insure that the conversion will be accomplished with a minimal loss of objectivity, low inference categories (Gage, N., 1969) are to be preferred. If two different classifications of each event are desired, a multiple coding technique (Galagher, 1960) can be used, in which two or more observers are coding simultaneously.

The answer to the second question depends on the objectives of the evaluation. This study is restricted to the instrumental and cognitive domains of classroom behavior. We confined our evaluation to these limits for two reasons: 1)we did not succeed in breaking the affective domain to low inference categories, and 2) affective behaviors are very rare in the microlab, especially in science lessons.

The main contribution of this study is a consequence of answering the last question. The first step toward evaluation was to state explicitly our values as to what is a "good" lesson. This is a question which has seldom been answered by any educational evaluator in an operative way, although it has been the subject of many books. The reason this question is so carefully avoided is an outcome of the fear of being superficial. In this study we decided to take the risk and state our values explicitly since we believe that evaluation is not an end in itself but a mean for achieving other goals. There is no harm in using values as long as they are explicit and stated beforehand. It can only enhance accomplishment of the training goals.
Two training objectives were defined:

1. increasing pupil participation and especially pupil initiation.
2. increasing the proportion of lesson time devoted to analytical and creative thinking.

The two sets of instrumental and cognitive categories for observation were arranged according to increasing pupil stimulation both instrumentally and intellectually. This hypothesized structure, after being verified empirically, served as a basis for giving weights to different classroom activities as will be explained later in the article. It also enabled calculation of a general score which reflects the achievement of the stated goals.

The objectives of this study were as follows:

1. To check the hypothesized structure empirically and to use it to construct the General Score (GS).
2. To investigate the General Score (GS) characteristics in a large population.
3. To study the distribution of the General Score (GS) as a result of different treatments.
4. To study the improvement in student-teacher teaching behavior as a result of written instructions as to how the GS is calculated.
5. To compare the change mentioned in 4. to a change which was caused by instructions with an unordered category observation system (FIAC).
6. To compare the plan of the lesson, prepared by the student-teacher in order to get the highest GS and the actual lesson conducted by him.

THEORETICAL FRAMEWORK

One of the most important developments in education in recent years is the Competency Based Teacher Education (CBTE) The following requirements of competency based instruction have been listed by Houston and Howsman: specifying behavioral objective, explicit criteria for determination whether performance met the indicated criterion level, public sharing of objectives, criteria and means of assessment, and placement on the learner of the accountability for meeting the criteria. Much has been written on the differences between CBTE and traditional training methods (Andrews, 1972, Daniel, 1971, Rosner, 1972)

The problem was whether certification should be based on teacher performance or if the results of that performance induce that pupil learning is the ultimate criterion, and thus the one one to be applied (Schalock, 1971). Others note that the teacher has little control over many pupil attributes and other factors which may influence learning to a greater extent than the teacher; thus the teacher can be held accountable only for his own behavior Popham (1973) is convinced that the only way to evaluate performance in a teaching performance test is by administering achievement tests to the learners He recommended that "clinical observers should conduct instructional analysis based on learner performance". However, he warns the reader to take care that it will not turn into a process-focused analysis. In this study we intend to evaluate the competency of a teacher in a teaching performance test according to his teaching behavior. The measurement is done by a category observation system, TDS, which is based on ordered facets. The order within the two facets has the same meaning and therefore the result is a partly ordered set. In the previous empirical testing a 2 dimensional space had shown up (Bar-On and Perlberg, 1973). This enables us to calculate one score for a lesson if we want a summative grade or two scores or more, if someone is interested in formative evaluation.

METHODS AND TECHNIQUES

The observation system consisted of twenty-four categories which are all the structuples (permutations) of Facets A and B:

Facet A - Teaching activities arranged according to the amount of pupil participation

a_1 - lectures verbally - teacher presents subject matter verbally without anticipating pupil response

a_2 - lectures non-verbally - teacher presents subject matter non-verbally without anticipating pupil response (shows diagrams, map, etc.)

a_3 - gives instructions - teacher asks pupil to carry out certain activity and dictates how this should be accomplished.

a_4 - asks questions - teacher solicits a reaction, without dictating its form.

a_5 - responds to pupil reaction - evaluates or criticizes pupil reaction or asks another question to clarify pupil reaction (probing questions)

a_6 - pupil response - pupil answers teacher's question without elaborating or adding any new idea.

a_7 - relates to pupil initiative - teacher encourages pupil to express his own ideas.

a_8 - pupil initiative - pupil asks teacher a question in order to gain more information or further elaboration. (If the pupil asks the teacher to repeat an explanation a_6 should be coded).

Facet B - Levels of thinking arranged according to the pupil's amount of intellectual stimulation.

b_1 - knowledge - teacher asks pupil a question related to memorized material or pupil answers a question by using previously mentioned information; also includes application of rules presented by the teacher previously e.g. mathematical example.

b_2 - analytical thinking - pupil infers and applies a rule differently than the way he was taught.

b_3 - creative thinking - pupil creates a new rule or suggests different solutions to the same problem.

It should be noted that the order in Facet A and in Facet B has the same direction of pupil stimulation. Therefore a partly ordered set was hypothesized:

INSERT FIG. 1 HERE

The assumption is that incomparable structuples get the same score. The hypothesized structure was checked again on the large sample of student-teachers for three years. The sample consisted of four random samples of four treatments. Each sample consisted of 32 student-teachers out of more than one hundred in each semester, so that the sample of 128 students can be considered as quite representative of the student-teacher population in the Technion.

Since our hypothesized structure was empirically confirmed on the large population (see chapter of "Results and Conclusion" - Fig. II), each category was weighted, as can be seen from Fig. I. The rationale for this weighting is that incomparable structuples get the same weight since we have not decided if higher thinking level with less pupil participation is better or worse than higher pupil participation on a lower level of thinking. As for comparable structuples, scales are formed, and therefore weights can be arbitrarily chosen (Guttman, L.A., 1944).

The structuple a_1b_1 was rated the lowest and given a grade of 1, the structuple a_8b_3 was rated highest and given a grade of 10. The other 22 structuples were rated according to their position in relation to the two grades mentioned above. The grade (GS) was calculated according to the following formula:

GS K_1 $K_2PT\sum X_iY_i$ where:

X_i - percentage of lesson time of each category i ($0 < X_i < 100$, i 1,2,...24)

Y_i - score of category i ($1 < Y_i < 10$)

P - proportion of pupils participating in the lesson (a participating pupil is one who takes part at least once in the lesson or a pupil whom the trainee tried to involve in the lesson) ($0 < P < 1$)

T - a number which depends on the number of transitions from one category to another (T will be calculated as 5 x number of transitions divided by 100, i.e. when the number of transitions is greater than twenty, T 1) ($0 < T < 1$)

K_1, K_2 - constants.

In this study K_1 and K_2 were set to 40.0 and 0.1 respectively. The function of these two constants is to obtain the desired range of GS. K_1 determines the mean of the GS distribution and K_2 determines its dispersion.

We can see from the formula for calculating GS that the more the student-teacher succeeds in bringing pupils to participate on an analytical and creative level of thinking the higher his grade will be. Since it's not possible to conduct an entire lesson in such a way that it will consist of pupil initiation on the analytical and creative thinking level, the two constants are needed.

P - The proportion of participating pupils is taken into account to make sure that there is not only one or two pupils who are speaking all the time.

T - Number of transition , was used to prevent the teacher from the possibility of getting a high GS by simply asking the pupils few high order questions which require long answers on a high thinking level without reacting to their answers. Therefore we insisted on having at least twenty transitions during the five minute lesson.

DATA SOURCE

There are two data sources in this study. One is a large sample of 128 student-teachers which consists of four random samples of 32 subjects. Each sample represents student-teachers who had different training methods. The large sample is used to investigate the hypothesized structure of intercorrelations among the observation categories. It is also used in investigating the characteristics of the GS. The second data source is sixty-five student-teachers, to whom the detailed written instruction about how the GS is calculated were given. They taught a lesson before and then got the instructions and were asked to plan a lesson to get the highest score possible. They planned their lessons on a time line display using the same categories. Then each student conducted his planned lesson in the microlaboratory. They were asked to plan a ten minute lesson but were allowed to choose the best five minutes. They were also told that the GS calculated from this performance test would serve as their grade for that semester. All lessons were videotaped and coded by trained observers. The coding of Facet A and Facet B was done simultaneously. A metronome was used and the observers were told to code the event that took place exactly when the metronome strike occurred. This was done to ensure that the observers are synchronized and refer to the same event.

All samples were drawn from the same population of student-teachers in the Teacher Training Department of the Technion. They were all in their second year of study and were studying to become science teachers in high school. The learners in the lessons were pupils from a junior high school. In all cases the lessons that served for the statistical analysis were "posttest" i.e. final lessons in which the trainee tried to do his best.

RESULTS AND CONCLUSIONS

Of the twenty-four possible categories only eight occurred sufficiently frequently to be used for statistical analysis. These eight categories represented both student participation and the level of thinking. To check the correspondence between the hypothesized structure and the empirical one, Smallest Space Analysis (SSA), was made. The computer program G-L SSA-1 (Guttman, L. 1967 , Lingoes, J.C., 1964) was used. This program finds a set of coordinates X_{ia} ($i=1,2,\ldots n$: $a=1,2,\ldots,m$) for the 8 categories such that if

$$d_{ij} = \sqrt{\sum_{a=1}^{m} (X_{ia} - X_{ia})^2}$$

then $d_{ij} \leq d_{kl}$ whenever $d_{is} \leq d_{kl}$ for m, a minimum number of Euclidean dimensions. The search is done after a distance function, d_{ij} is defined on the 8(8-1)1/2 pairs of points We did it by calculating the Pearson Product moment correlation for each one of the 8(8-1)1/2 possible pairs. The G-L(SSA-I) output contains a space diagram in which each variable is represented by a point in this space according to its pair of coordinates. The resultant two-dimensional diagram is given in Fig. II.

INSERT FIG. II HERE

When the hypothesized structure which served as the basis of calculating the GS, was confirmed empirically, we started to study the GS characteristics. A GS was computed for each of the pre- and post-lessons for each of the 128 student-teachers. A frequency distribution was drawn for the GS 50 - 100 range. The distribution looked reasonably normal for both pre- and post-lessons. We have not found a test for non-normality so we used a chi-square test for normality. The problem is that this test assumes the normality of a distribution as its null hypothesis. We did not succeed in rejecting this hyphothesis even for α as large as 0.85. This does not prove, of course that the distribution is normal because a null hypothesis can never be confirmed. The problem could be approached as a line fitting problem but this was not done in this study. The GS ranged from 50 - 100; with a mean of 74 and standard deviation of 9.6.

The 3rd objective was to study the distribution of the GS score in selected samples. A frequency distribution diagram was drawn for each one of the four samples each of which represents another treatment (Perlberg, Bar-On, Levin, and Etrog, 1974). The results were as follows:

1. In the group which was trained in a microteaching laboratory neither with the aid of an observation system (TDS) nor with the method of focusing on a specific skill, the distribution of GS was normal with the lowest mean ($M = 60$) and a small spread of scores.
2. In the group that was trained with both systems (Focusing on a Specific Skill and an ordered category observation system- TDS) and also was aware of how the GS is calculated, the distribution was also normal with a mean of 76 and a larger range of scores.
3. In the group that was trained only with the aid of feedback focused on a specific skill (like the microteaching method recommended by Allen and Ryan, 1969) the distribution was bimodal with a mean of 72 and a relatively large range.
4. In the group that was trained with the aid of an ordered category observation system only the distribution was also bimodal with the highest mean ($M = 90$) and the largest dispersion.

It is too soon to draw any conclusions but it seems that the distribution of GS is normal when there is no training at all (pre-lessons) or when the training is with several methods so that each student-trainer can find the methods that best fits his needs. In training methods which use only one feedback system in which the trainees are not aware of how the GS is calculated, the distribution is more dispersed and multimodel because this training method fits only the needs of some trainees and is not as adequate for others,

The fourth objective was to investigate the change in student-teacher teaching performance as measured by the GS, using written instruction alone. We did this experiment at the end of the semester with students that were already trained with the aid of TDS feedback and feedback which was focused on a specific skill. The posttest, which served also to determine their semester grades, was compared to the previous lesson before the written instructio were handed out. There was no significant difference between the arithmetic means of the two lessons. The main effect was that the GS scores in the post-lesson were less spread out and the distribution became normal. This indicates that the wirtten instructions are of more help to those students who had low GS grades. When this experiment was done with untrained teachers a very dramatic

change occurred but we do not yet have the final results. It is known that instructions alone have a very great affect on modifying teacher performance, because part of the improvement in teaching performance is simply a result of the student-teacher's awareness of what is expected of him. The comparison between the improvement due to written instructions about unordered category system (FIAC) and an ordered one (TDS) yields some interesting results. When comparing improvement as a result of instructions (FIAC) with the change as is caused by training (Klinzing, Bar-On 1974) it was found that some changes such as decreasing "Teacher Talk Ratio" can be achieved by instructions alone, while improving a category such as "Teacher Uses Pupil Ideas" needs training.

Written instructions on the way the GS is calculated effect a change in the proportions of the various activities in such a way that the General Score is maximized. Therefore, the difference in the outcome is that in an unordered system (FIAC) the improvement is usually in one category while in an ordered system it is in all categories. The student-teacher who plans a lesson which will maximize his GS starts from the end: in order to get a high score most of the lesson must be pupil initiations and respones in a high level of thinking. For the purpose of evoking pupil participation which involves high thinking level he must ask analytical and divergent questions. But for asking analytical questions he usually needs to explain something or to lecture. For each three seconds of lecturing he gets a low grade. Questions get a higher score but much lower than the score for pupil response or initiation. So, the teacher has to decide what is the minimal investment of time which is worth spending on low score activities which will serve as the basis for stimulating pupil participation.

Each student handed in his lesson plan before taking the performance test. Since in the lesson plan the same categories were used, it was possible to calculate the GS for the planned lesson and it compare it to the GS of the actual lesson. There was a negative low correlation between the two scores ($r=-0.26$). When the students were divided into two groups according to their GS it was found that students with the higher grades were much more realistic in their planning, while the low grade trainees were very unrealistic and planned a lesson which consisted of pupil initiation at a creative level without any teacher stimulated activities, e.g.: trainees that got an actual GS = 65 planned a lesson that was evaluated as high as 135, while trainees who scored an actual GS = 85 planned a lesson rated 90.

Comparison of time lines between planned and actual lessons yields that most of them look very different. When a transition probability matrix was calculated for each planned lesson and then correlated with the actual lesson matrix the correlation was low. The only thing that remained consistent from planned to actual lesson was the structure of the intercorrelation matrix. A Smallest Space Analysis was done on the planned lesson as can be seen in fig. III.

INSERT Fig. III HERE

A typical Porex structure can be seen from the two dimensional space diagram. All the analytical activities form a simplex structure while the knowledge activities form a parallel simplex. The order of structuples is also typical to that of performance tests: the category "Teacher Asks" is considered higher than "Teacher Reacts To Pupil Answer", because under the pressure of test situations a teacher finds asking questions a better investment than relating to pupil answers. The same structure for test lessons was found in previous studies (Perlberg et al, 1973). Since in the planned lessons, creative activities were also present, they form another simplex.

Smallest space analysis was made also on both the planned and actual lesson. It can be clearly seen that the structure remains the same (see fig. IV).

INSERT Fig. IV HERE

The last result was that the GS was highly correlated (r 0.74) with the supervisor's general evaluation of the student teaching performance. Most of the students preferred the computer evaluation because they found it more objective, and even students with low grades felt that they can blame no one but themselves.

FACET A	Yi	
a_1-teacher lectures verbally	1	a_1b_1
a_2-teacher lectures non-verbally	2	a_2b_1 a_1b_2
a_3-teacher gives directions	3	a_3b_1 a_2b_2 a_1b_3
a_4-teacher asks	4	a_4b_1 a_3b_2 a_2b_3
$a_5$teacher reacts to pupil response	5	a_5b_1 a_4b_2 a_3b_3
a_6-pupil answers	6	a_6b_1 a_5b_2 a_4b_3
a_7-teacher reacts to pupil initiative	7	a_7b_1 a_6b_2 a_4b_3
a_8-pupil initiates	8	a_8b_1 a_7b_2 a_6b_3
FACET B	9	a_8b_2 a_7b_3
b_1-knowledge	10	a_8b_3
b_2-analytical		
b_3-creative		

Fig 1-The partly ordered set which results from the Cartesian product of the two ordered facets.Moving from top to bottom denots increas in p. stimulation.

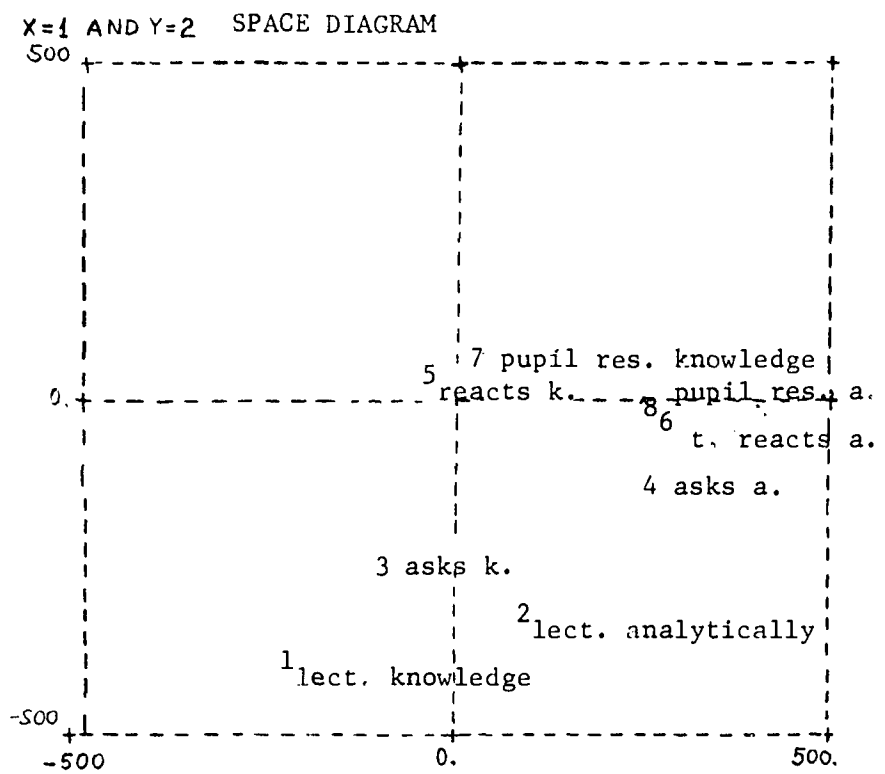

Fig 2- Space diagram of 128 student-teachers reveals the empirical structure.

X= 1 AND Y = 2 SPACE DIAGRAM

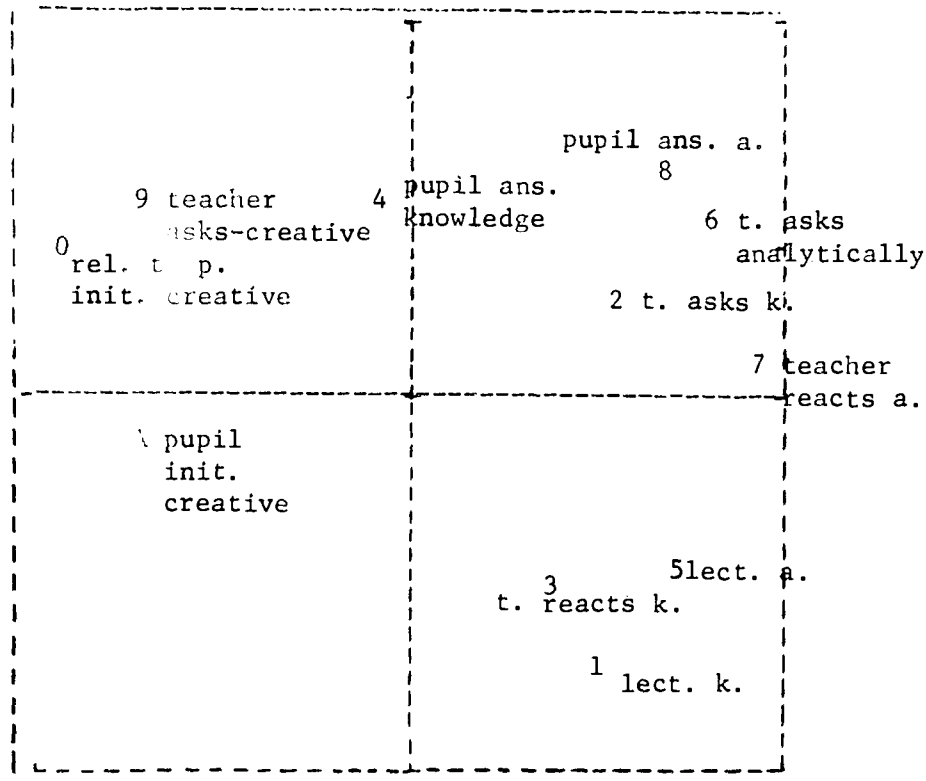

Fig. 3 - Space diagram of the planned lesson.

X 1 AND Y 2 SPACE DIAGRAM

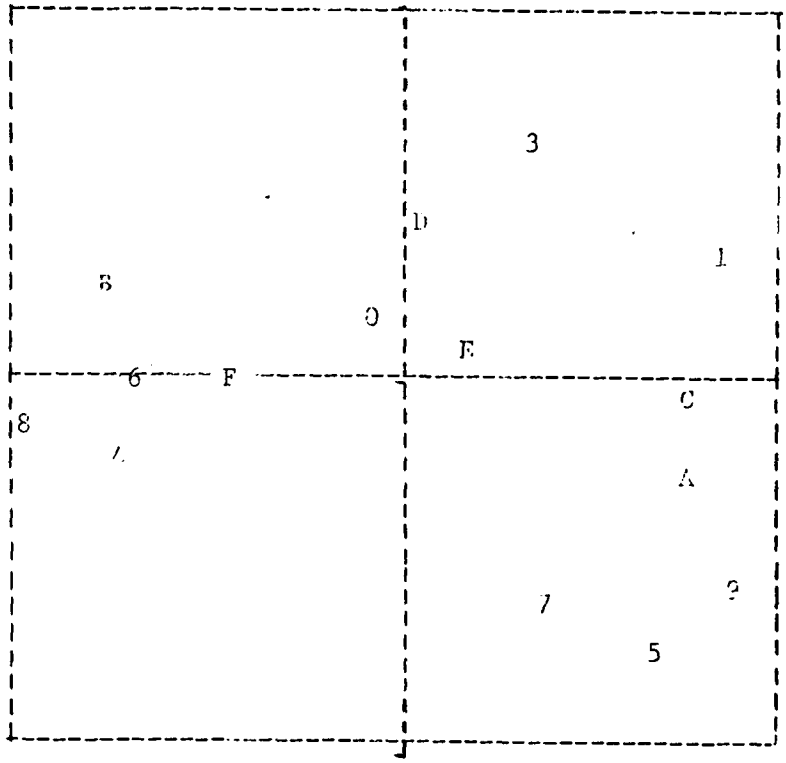

Fig. 4-Space diagram Of the planned and actual lesscn. Categories 1-8 belong to the actual lesson,categories 9-F belong to the planned lesson. Categories 1,3,..,E. are knowledge,Categories 2,4,..,F.are analytical.